# Biomedical Named Entity Recognition with Tri-training learning

YueHong Cai
Foreign Language Learning Center
JIANGSU University
Zhengjiang, China
caiyh@ujs.edu.cn

XianYi Cheng
School of Computer Science &Communication Engineering
JIANGSU University
Zhengjiang, China
xycheng@ujs.edu.cn

***Abstract*—In order to solve the data scarcity problem, this paper presented a co-training style method for Biomedical Named Entity Recognition. We proposed a novel selection method for tri-training learning, using three classifiers: CRFs,SVMs and ME. In tri-training process, we select new newly labeled samples based on the selection model maximizing training utility, and compute the agreement according to the agreement scoring function. Experiments on GENIA corpus show that our proposed tri-training learning approach can more effectively and stably exploit unlabeled data to improve the generalization ability than Co-training and the standard Tri-training.***

***Keywords- biomedical named entity recognition;semi-supervised learning; tri-training***

## I. INTRODUCTION

With the explosion of information in the biomedical domain, the wealth of biomedical knowledge in the form of semi-structured documents in various text documents. Therefore, biomedical text mining and biomedical information extraction have received increased attention in recent years. Biomedical named entity recognition (Bio-NER) is one of the most elementary and core tasks in the biomedical knowledge discovery.

The Bio-NER problem is about how to extract the biomedical named entities (NEs) of interest (such as PROTEIN, DNA, RNA etc.) from biomedical literature. There are three main approaches to Bio-NER, rule-based approach [1], dictionary-based approach [2], and machine learning approach. Currently, machine learning approaches have become increasingly dominant Bio-NER technology, there are some research efforts using machine learning techniques to recognize biomedical NEs in texts. These techniques include Hidden Markov Model (HMM) [3], Support Vector Machines (SVMs) [4], Maximum Entropy Markov Model (MEMM) [5], Conditional Random Fields (CRFs) [6],[ 7], etc.

To improve the generalization ability, the amount of manually labeled training data required by supervised learning methods is quite large. However, the limited availability of labeled data in most domains, including the biomedical, restricts the application of such methods. Semi-supervised learning that exploits unlabeled data, in addition to labeled ones, has recently become an active research area.

In this paper, we investigate the use of a semi-supervised learning approach, tri-training learning[8], on Bio-NER .By considering that Bio-NER is a sequence labeling problem, we propose a novel approach of selecting training samples for tri-training learning. The experimental on GENIA corpus show that the proposed approach can improve the performance significantly using a large pool of unlabeled data.

## II. TRI-TRAINING FOR BIO-NER

Tri-training is motivated from co-training style approach, which uses three classifiers of the same algorithm. Tri-training bootstrap-samples the original labeled data to generate three different training sets, then each classifier is initially trained from a data set. In the co-training style process of Tri-training, for any classifier, an unlabeled example can be labeled for it as long as the other two classifiers agree on the labeling of this example, while the confidence of the labeling of the classifiers are not needed to be explicitly measured. The generation of the initial classifiers looks like training an ensemble from the labeled example set with a popular ensemble learning algorithm. Tri-training algorithm does not require sufficient and redundant views and the constraints on the employed classifiers, nor does it require time-consuming cross validation process. Therefore, it has more applications and more efficient.

In order to keep classifiers diverse, we propose to use more than one classifier with different algorithms. In this paper, we conducted experiments with three different learning algorithms: CRFs, SVMs and Maximum Entropy models (ME).

### *A. Learning algorithms*

**CRFs:** CRFs are probabilistic frameworks for sequence segmentation and labeling tasks, which has been successfully applied to many natural language

978-1-4244-4134-1/09/$25.00 ©2009 IEEE

processing tasks. They are undirected graphical models that provide the conditional probability distribution of a label sequence $y=y_1, y_2, \cdots, y_n$ given the observation sequence $x=x_1, x_2, \cdots, x_n$. A linear-chain CRF defines a conditional probability as follows:

$$p(y \mid x) = \frac{1}{Z_o} \exp \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x, i) \qquad (1)$$

Where $Z_o$ is the normalization factor, $f_j(y_{i-1}, y_i, x, i)$ is a binary-valued feature function and $\lambda_i$ is its weight.

**SVMs:** SVMs have excellent classification ability and generalization ability which based on the structural risk minimization of statistical learning theory. The basic approach adopted by SVMs is to construct a simple binary classification function by mapping the input patterns to a higher dimensional feature space if the patterns cannot be linearly separated in input space and to seek an optimal separating hyper-plane to divide the training instances into two classes. The mapping function uses a dot product of the input space pattern vectors transformed according to a kernel function. In this paper, we apply "pair wise method" to construct multi-class SVMs by combining several binary SVMs.

**ME:** ME is a powerful method for constructing statistical models of classification tasks, which has been successfully employed for many natural language processing tasks. The principle of maximum entropy offers a clean way to choose probability distribution that satisfies the above property is the one with the highest entropy subject to the constraints imposed (Such constraints are expressing some relationship between features and outcome). The idea is to be "maximally noncommittal" about what we do not know, while still agreeing with what we do know. The advantage of ME is that it is an extremely flexible technique, since it can use a virtually unrestricted and rich feature set in the framework of a probability model.

### *B. Tri-training algorithms*

In [8], the detail of Tri-training algorithm is studied. In this paper, we give two proposals to improve the standard Tri-training algorithm. First, the final joint classifier $\{H_1, H_2, H_3\}$ classifies a new data via accuracy weighted voting principle according to Equation 2, which employs accuracies on initial seeds as weights. Second, in each round, a new sample is selected based on the agreement measure.

$$H(1,2,3)(x) = \arg \max_{y \in label} \frac{\sum_{i=1}^{3} \delta(y, H_i(x)) \times P_i(L)}{\sum_{i=1}^{3} P_i(L)} \qquad (2)$$

Where $\delta(y, H_i(x)) = \begin{cases} 1 & H_i(x) = y \\ 0 & H_i(x) \neq y \end{cases}$

The pseudo-code of the algorithm of tri-training for Bio-NER is presented in Figure 1. At each iteration, a cache of samples is selected from the total pool of unlabelled samples.

$H_1$、$H_2$ and $H_3$ are three different classifiers.
$M_1^i$, $M_2^i$ and $M_3^i$ are the models for $H_1$、$H_2$ and $H_3$ at step *i*.
U: the original unlabeled data, L: the original labeled data.
$U^i$ is a small cache holding a subset of U at step *i*.
$L_1^i, L_2^i$ and $L_3^i$ are the labeled data for $H_1$,$H_2$ and $H_3$ at step i.

**Initialize:**

$L_1^0 \leftarrow L$, $L_2^0 \leftarrow L$, $L_3^0 \leftarrow L$, $S \leftarrow L$

$M_1^0 \leftarrow Train(H_1, L_1^0)$,

$M_2^0 \leftarrow Train(H_2, L_2^0)$,

$M_3^0 \leftarrow Train(H_3, L_3^0)$

**Loop:**

$L_1^i \leftarrow BootstrapSample(S)$,

$L_2^i \leftarrow BootstrapSample(S)$,

$L_3^i \leftarrow BootstrapSample(S)$

$U^i \leftarrow$ Add unlabeled data from U.

$M_1^i$, $M_2^i$ and $M_3^i$ tag the data in $U^i$ and compute scores of agreement between them according to the agreement functions.

Select newly label data $\{P_1\}$、$\{P_2\}$ and $\{P_3\}$ according to some selection method.

$L_1^{i+1} \leftarrow L_1^i + \{P_1\}$, $L_2^{i+1} \leftarrow L_2^i + \{P_2\}$, $L_3^{i+1} \leftarrow L_3^i + \{P_3\}$

$M_1^{i+1} \leftarrow Train(H_1, L_1^{i+1})$,

$M_2^{i+1} \leftarrow Train(H_2, L_2^{i+1})$,

$M_3^{i+1} \leftarrow Train(H_3, L_3^{i+1})$

$S \leftarrow L_1^{i+1} + L_2^{i+1} + L_3^{i+1}$

$M_1^{i+1}$, $M_2^{i+1}$ and $M_3^{i+1}$ tag the data in S based on the weighted strategy for Ensembles of Classifiers

Figure 1. The algorithm of tri-training for Bio-NER

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

### *C. Sample Selection Metric*

In this paper, the problem of Bio-NER can be regarded as a sequence labeling task. We select new newly labeled samples according to the selection model maximizing training utility, and compute the agreement based on the agreement scoring function.

*1) Agreement scoring function:*Given a data sequence $X=\{x_1, x_2, \ldots x_n\}$,determine two labeled sequences $Y_1 = \{y_{11},y_{21},\ldots,y_{n1}\}$,$Y_2 = \{y_{12},y_{22},\ldots,y_{n2}\}$ using two classifiers. Now, the agreement scoring function *f(Ag)* is defined as follows:

$$f(Ag) = \sum_{i=1}^{n} f(y_{i1}, y_{i2}) / n \qquad (3)$$

Where $f(y_{i1}, y_{i2}) = \begin{cases} 1 & y_{i1} = y_{i2} \\ 0 & y_{i1} \neq y_{i2} \end{cases}$ $\psi$

The agreement scoring function denotes the agreement between two labeled sequences or the agreement between two classifiers. The larger *f(Ag)* is, the higher the agreement is.

*2) Sample Selection Method:*Let three different classifiers $H_1$, $H_2$ and $H_3$, $x$ is a data in unlabeled data. As [9] suggested, we should select new samples, which have high training utility. Selecting principles are shown below:

1. If the higher agreement scores between the classifiers $H_2$ and $H_3$ at an example $x$, then $x$ can be correctly labeled for $H_1$.

2. If the classifier $H_1$ disagree with the other classifiers ($H_2$ and $H_3$) at an example $x$, then $x$ is incorrectly labeled for $H_1$.

Let $C_i$ *(i=1, 2, 3)* denote the newly labeled training subset by $H_i$ *(i=1, 2, 3)* from the cache. By applying two above-mentioned principles, we propose the selection model maximizing training utility. In each round of tri-training process, selection is performed as follows:

1. Compute the agreement scores for all samples in $C_j$ and $C_k$ *(j, k ≠i),* and then choose the subset of $n$.percent-highest scoring samples by $C_j$ and $C_k$;

2.Compute the agreement scores for all samples in $C_i$ *and* $C_j$, and then choose the subset of $n$. percent-lowest scoring labeled samples by $C_i$ and $C_j$;

3.Apply the intersection. Then the new samples that are selected the intersection are labeled by $H_j$.

The selection method's control parameter decides the number of labeled samples to add at each iteration. It also acts as an indirect control of the number of errors added to the training set. A strict control parameter is very important. In our experiments, we set the selection method's control parameter as 30% by referencing to [9].

## III. EXPERIMENTS AND DISCUSSION

### A. Data set and Evaluation measures

The labeled data is from GENIA 3.02 corpus, which contains 2,000 abstracts. It has been annotated with semantic information (such as PROTEIN, DNA) and POS information. In the experiments, we divided the GENIA corpus into two parts: 800 abstracts were used as the original labeled set, and the other 1200 abstracts were used as testing set. The unlabeled data is downloaded form MEDLINE, which contains 487,458 abstracts (6G XML abstract data).

The experimental results are evaluated by F-scores using JNLPBA's evaluation script, which is a modified version of the CoNLL evaluation script. The F-score is defined as follows:

$$F = (2PR) / (P + R) \qquad (4)$$

Where $P$ denotes Precision and $R$ denotes Recall.

### B. Feature selection

Feature selection is essential to any statistical machine learning models. Our system uses six singleton feature types: word, orthographical, part-of-speech (POS), word class, affix, and chunk.

**Word features:** the word features include unigram, bigram and trigram.

**Orthographic features**: the orthographic features include capital letter, dash, punctuation, and word length.

**Word Class**: WC features replace a capital letter with "A", a lower case letter with "a", a digit with "0", and all other characters with "_". Similar brief word class (BWC) features are added by collapsing all consecutive identical characters in word class features into one character. For example, for the word Kappa-B, WC = Aaaaa A , and BWC = Aa A.

**POS features**: the POS features are provided by the GENIA tagger.

**Prefix and Suffix Features:** Some prefixes and suffixes can provide good clues for classifying named entities. For example, words which end in "ase" are usually proteins. In our experience, the acceptable length for prefixes and suffixes is 3-5 characters.

**Chunk features**: This feature is also effective in determining the position of a word in NEs, "B", "I", "O" which respectively means "begin chunk", "in chunk" and "others".

### C. Results and Analyses

In order to investigated the performance of different selection methods. In our experiment, the performance of the selection model maximizing training utility is compared with the selection model in [8].In contrast to tri-training, we also investigate the performance of co-training to Bio-NER. In co-training process, a subset of the newly labeled samples is selected by the agreement scoring function.

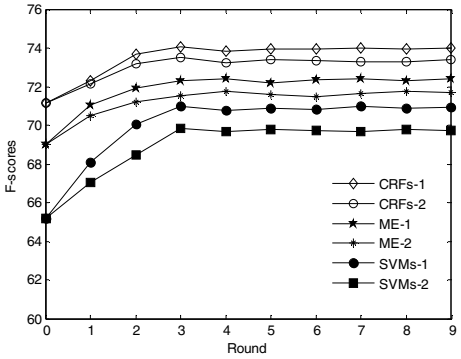


Figure 2. Results of different selection methods

Figure 2 shows the F-score rates of different selection methods, where CRFs-1 refers to the CRFs model with the selection model maximizing training utility, CRFs-2 refers to the CRFs model with the standard Tri-training and the other strings have similar meanings. From the figure, we found that the final hypotheses generated are apparently better than the initial hypotheses with all the classifiers in all situations. Moreover, it is obvious that the performance of Tri-training based on the selection model maximizing training utility significantly outperforms the standard Tri-training. This proves that our proposed tri-training can effectively exploit unlabeled examples to enhance the learning performance.

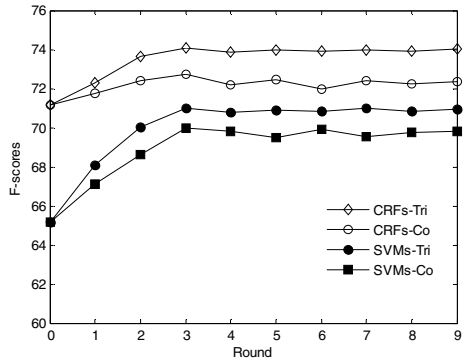


Figure 3. Tri-training vs Co-training

The F-score rates of the compared algorithms are depicted in Figure 3. Compared with co-training, our proposed tri-training learning approach achieved better performance at each iteration. Figure 3 reveal that the Tri-training algorithm is more stable than the Co-training in performance, and scarcely appears the fluctuating phenomenon in the process. Experimental results show that the proposed method can reduce the error rates of the rounds and improve both the generalization ability and the robustness.

## IV. CONCLUSION

This paper presented a new co-training style semi-supervised learning algorithm for Biomedical Named Entity Recognition. This algorithm uses three different classifiers (CRFs、SVMs and MEMM) to exploit unlabeled data in the tri-training process. In each iteration, a new sample is selected by considering the agreements of three classifiers. The experimental results showed that the proposed approach can effectively exploit unlabeled data to enhance generalization ability

We presented an experimental study in the data scarcity problem of Bio-NER.In our experiment, the fraction of original labeled training set for Tri-training is too small to train an individual classifier with high accuracy. More and more inevitable noises due to misclassification are generated during the Tri-training process. How to effectively identify the wrongly labeled examples and filter noise is an interesting issue to be investigated to enhance the classification performance of co-training style approaches in future work.

## References


[1] F. Olsson, G. Eriksson, and K. Franzen, L. Asker, and P. Liden, "Notions of correctness when evaluating protein name taggers," Proceedings of the 19th international conference on computational linguistics, Taipei, Taiwan; 2002. pp. 765–71.

[2] Z.H. Yang, H.F. Lin, and Y.P. Li,"Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature," Computational Biology and Chemistry, vol.32,pp.287-291,2008

[3] J. Zhang, D. Shen, G.D. Zhou, J. Su, and C.L. Tan, "Enhancing HMM-based biomedical named entity recognition by studying special phenomena," Journal of Biomedical Informatics, vol.37,pp.411-422,2004

[4] T.Koichi,and C.Nigel,"Bio-Medical Entity Extraction using Support Vector Machines," Artificial Intelligence in Medicine, vol.33,pp.125-137,2005

[5] Finkel J, Dingare S, Nguyen H, Nissim M, Manning C, and Sinclair G, "Exploiting context for biomedical entity recognition: from syntax to the web," Proceedings of the joint workshop on natural language processing in biomedicine and its applications, Geneva, Switzerland;2004. pp. 88–91.

[6] T.H. Tsai , W.C. Chou , S.H. Wu,T.Y. Sung, J. Hsiang and W.L. Hsu, "Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities," Expert Systems with Applications,vol.2006,pp.117-128,2006

[7] C.J. Sun, Y. Guan, X.L. Wang, and L. Lin ,"Rich features based Conditional Random Fields for biological named entities recognition," Computers in Biology and Medicine, vol.37,pp.1327-1333,2007

[8] Z.H. Zhou, and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," IEEE Transactions on Knowledge and Data Engineering ,vol.17,pp. 1529–1541,2005

[9] M. Steedman,R. Hwa,S. Clark,M. Osborne,A. Sarkar,J. Hockenmaier_,et al,"Example selection for bootstrapping statistical parsers," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,Edmonton,2003,pp.157–164